Complete Suboptimal Folding of RNA and the Stability of Secondary Structures

Stefan Wuchty[1], **Walter Fontana**[1,2], **Ivo L. Hofacker**[1], **and Peter Schuster**[1,2]

[1]Institut für Theoretische Chemie
Universität Wien
Währingerstrasse 17
A-1090 Wien, Austria

[2]Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87501 USA

{steve,ivo,pks}@tbi.univie.ac.at
walter@santafe.edu

SYNOPSIS

An algorithm is presented for generating rigorously all suboptimal secondary structures between the minimum free energy and an arbitrary upper limit. The algorithm performs particularly well in the vicinity of the minimum free energy. This enables the efficient approximation of statistical quantities, such as the partition function or measures for structural diversity. The density of states at low energies and its associated structures are crucial in assessing from a thermodynamic point of view how well defined the ground state is. We demonstrate this by exploring the role of base modification in tRNA secondary structures, both at the level of individual sequences from *E. coli* and by comparing artificially generated ensembles of modified and unmodified sequences with the same tRNA structure. The two major conclusions are that (1) base modification considerably sharpens the definition of the ground state structure by constraining energetically adjacent structures to be similar to the ground state, and (2) sequences whose ground state structure is thermodynamically well defined show a significant tendency to buffer single point mutations. This can have evolutionary implications, since selection pressure to improve the definition of ground states with biological function may result in increased neutrality.

INTRODUCTION

The structure of RNA molecules can be discussed at an empirically well established level of resolution known as secondary structure. It refers to a topology of binary contacts arising from specific base pairing, rather than a geometry cast in terms of coordinates and distances (see Figure 1). The driving force behind secondary structure formation is the stacking of base pairs. The formation of an energetically favorable helical region, however, also implies the formation of an energetically unfavorable loop region. This "frustrated" energetics leads to a vast combinatorics of helix and loop arrangements spanning the structural repertoire of an individual RNA sequence.

The secondary structure provides both geometrically and thermodynamically a scaffold for the tertiary structure. Its free energy accounts for a large share of the overall free energy of the full structure. This linkage puts the secondary structure in correspondence with functional properties of the tertiary structure. Consequently, selection pressures become manifest at the secondary structure level as evolutionary conserved base pairs.

A secondary structure can be conveniently discretized as a graph representing a pattern of base pair contacts (Figure 1). This yields a formally well-defined combinatorial object which can be subject to mathematical treatment. Of particular interest are secondary structures possessing some extremal property with respect to a given sequence, such as having the largest number of admissible base pairs or minimizing the free energy. The theoretical importance of RNA as a model system for sequence-structure relations in biopolymers lies in the fact that structures of this kind can be computed by dynamic programming [1, 2, 3, 4, 5]. This method produces a single structure with the desired extremal property (even in the case of degeneracy). It has been stressed [6], however, that this may not adequately describe a real situation for two major reasons. First, the energy parameters on which the folding algorithm relies are inevitably imprecise. Hence, the true minimum free energy (mfe) structure might be one that is suboptimal with respect to the parameters used. The same might hold because of unknown biological constraints which may change relative energies, turning an otherwise suboptimal structure into the most favorable one. Second, under physiological conditions RNA sequences may exist in alternative states whose energy difference is small. Aside from their possible biological significance, the density and accessibility of such low lying states may determine how well-defined an mfe structure actually is.

Issues like these have prompted several approaches to generating suboptimal structures [6, 7, 8]. While representing an improvement, these approaches share one problem: they don't compute **all** suboptimal structures within a given energy range from the mfe. For example, a widely used algorithm is Zuker's extension [6] of his own dynamic programming procedure [4]. It generates for each admissible base pair in a given sequence the energetically best structure containing that base pair. Hence, for a sequence of length n at most $n(n-1)/2$ suboptimal structures are produced. Furthermore, each base pair present in the mfe structure regenerates by definition the mfe structure as the best structure containing it. It follows that no structures are generated which differ from the mfe by the absence of one or more base pairs. In addition, if an mfe structure consists of two substructures A-B connected by a stretch of external bases, no suboptimal alternatives will be produced that are suboptimal in both modules. As a calibration for the number of structures missed, consider the *E.coli* sequences $\mathrm{tRNA}^{\mathrm{his}}$ (RH1660) and $\mathrm{tRNA}^{\mathrm{ser}}$ (RS1661) from the EMBL Heidelberg tRNA Database (see Appendix C) which have 5 and 73 structures, respectively, within 10% of their minimum free energy. Of these, 2 ($\mathrm{tRNA}^{\mathrm{his}}$) and 17 ($\mathrm{tRNA}^{\mathrm{ser}}$) structures would show up under Zuker's scheme.

Many of the missing structures may well be classified as "uninteresting" by some account, yet this cannot be said with certainty for all of them and not for all accounts. They clearly are relevant in the calculation of measures for structural well-definedness, in approximating statistical quantities such as the partition function, or in calculating the density of states at low energies. Among the major benefits of a complete suboptimal folding procedure is the possibility of rigorously analyzing the low energy section of the energy landscape on which the actual kinetic folding process occurs.

In this paper we describe a fairly simple algorithm which generates all suboptimal folds of a sequence within a desired energy range from the mfe. The idea underlying the algorithm is straightforward, and we took it literally from Waterman and Byers [9] who developed it in the context of suboptimal solutions to the shortest path problem in networks. Waterman and Byers also applied their scheme to obtain near-optimal sequence alignments [9]. Yet, to our knowledge, their idea has not been exploited to produce suboptimal solutions to RNA folding, which is somewhat puzzling, since the energy minimization of secondary structures is handled by the same technique employed in the shortest path problem or in sequence alignment. We first illustrate the

Waterman-Byers scheme for the case of base pair maximization. While being of theoretical interest only, the case serves as a pedagogical exposition of the logic underlying the algorithm. We then briefly discuss the more involved case of energy minimization, while relegating excruciating details to appendices A and B. We proceed by applying the algorithm to study the degree to which a minimum free energy structure is thermodynamically sharply defined. We are specifically interested in the role of base modification in tRNA sequences to that effect.

MAXIMUM MATCHING AND THE WATERMAN-BYERS SCHEME

The usual formalization [5, 10] views a secondary structure as a graph whose nodes represent nucleotides at positions $i = 1, \dots, n$ of an RNA sequence of length n. The set of edges connecting the nodes consists of two disjoint subsets. One is common to all secondary structure graphs, while the other is specific to each sequence. The common set represents the covalent backbone connecting node i with node $i+1$, $i = 1, \dots, n-1$. The sequence specific part consists of a set P of edges $i \cdot j$, $\mathcal{P} = \{\ i \cdot j \mid i \neq j \text{ and } j \neq i+1\ \}$, representing admissible hydrogen bonds between the bases at positions i and j, such that (i) every edge in $\mathcal{P}$ connects a node to at most one other node, and (ii) the pseudoknot constraint is met. The latter states that if both $i \cdot j$ and $k \cdot l$ are in $\mathcal{P}$, then $i < k < j$ implies that $i < l < j$. Failure to meet this constraint results in interactions which are considered to be tertiary (pseudoknots), or, perhaps more to the point, computationally and thermodynamically unwieldy at present. The set of admissible base pairs which we shall consider consists of the Watson-Crick pairs {**AU,UA,GC,CG**} and {**GU,UG**}.

The problem of finding the largest possible set $\mathcal{P}$ of admissible base pairs compatible with the above definition of a secondary structure is known as "maximum matching". A **matching** in an undirected graph G is a set of edges, no two of which have a vertex in common. Evidently, any set $\mathcal{P}$ of base pairs compliant with the definition of secondary structure is a matching. A matching M is called a maximum matching, if no matching contains more edges than M.

When maximizing base pairing, the basic structural building block is an

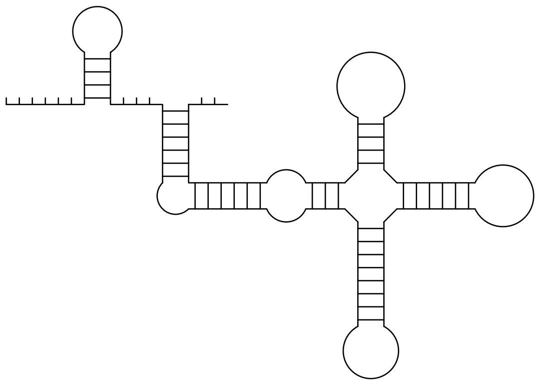


Figure 1: **An RNA secondary structure graph.** Unpaired positions not enclosed by base pairs, such as free ends or links between independent structure modules, are called "external". Here they are marked by ticks.

individual base pair. This is in contrast to energy minimization, where the building blocks to which an energy can be assigned are larger chunks of context known as "loops" (or faces of the secondary structure graph). It is this property which makes maximum matching considerably simpler than free energy folding.

The dynamic programming procedure to compute the maximum number of admissible base pairs is straightforward [1, 10]. Let $P_{i,j}$, $i < j$, denote the maximum number of base pairs on the sequence segment $[i, j]$. $P_{i,j}$ can be defined recursively:

$$P_{i,j} = \max \left\{ P_{i,j-1}, \max_{i \leq l \leq j-2} \left\{ \left(P_{i,l-1} + 1 + P_{l+1,j-1} \right) \rho(a_l, a_j) \right\} \right\} \qquad (1)$$

where $a_i \in \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$ denotes the base at position i, and $\rho(\cdot, \cdot)$ is an indicator function of biophysically legal base pairs:

$$\rho(a_i, a_j) = \begin{cases} 1 & \text{if } a_i \text{ and } a_j \text{ can pair,} \\ 0 & \text{otherwise.} \end{cases}$$

The recursion in equation (1) works by filling the P array in such a way that all smaller fragments needed in the computation of $P_{i,j}$ have already

$\mathcal{S} = \{\sigma = \{[1, n]\}; \mathcal{P} = \emptyset\}$
repeat:
if $(\sigma = \emptyset)$
    terminate with $\mathcal{S} = \{\emptyset; \mathcal{P}\}$
$[a, b] \Leftarrow \sigma$
if $(P_{a,b} = P_{a,b-1})$
    $[a, b-1] \Rightarrow \sigma$ and repeat
if ($l \cdot b$ with $l \in [a, b-2]$ and $P_{a,b} = P_{a,l-1} + P_{l+1,b-1} + 1$)
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{l \cdot b\}, [a, l-1] \Rightarrow \sigma, [l+1, b-1] \Rightarrow \sigma$ and repeat

Table 1: **Pseudocode for the backtrack process in the maximum matching problem.** $x \Leftarrow \sigma$ denotes the popping of the first element from the stack σ. That element is assigned to x, and is deleted from σ. $x \Rightarrow \sigma$ denotes the pushing of x on the stack σ.

been computed. (For example, let index i run from n down to 1, while index j sweeps from $i+2$ to n.). Adding bases sequentially at the $3'$-end, the procedure (1) checks whether a pairing between the added base and some position downstream improves the total number of pairs on the segment, as compared to leaving the added base unpaired. When all is done, the maximum number of base pairs is $P_{\max} = P_{1,n}$. A structure with $P_{\max}$ pairs is obtained by tracing back through the P array. Although the backtrack is simple, we shall explain it in some detail, since it is the key procedure in understanding the Waterman-Byers extension.

Let us define a **partial structure** $\mathcal{S}$ to be a pair $\mathcal{S} = (\sigma; \mathcal{P})$ consisting of a stack σ of sequence segments $\{[i_1, i_2].[i_3, i_4]. \cdots\}$, and a set $\mathcal{P}$ of base pairs. A **complete structure** is a partial structure whose stack is empty, $\mathcal{S} = (\emptyset; \mathcal{P})$.

The backtrack starts with the partial structure $([1, n]; \emptyset)$, pops the segment from the stack, and, following equation (1), checks whether the nth position shall remain unpaired, i.e., whether $P_{1,n} = P_{1,n-1}$. In that case, $[1, n-1]$ is pushed on the stack σ, and the procedure repeats similarly with $\mathcal{S} = ([1, n-1]; \emptyset)$. If $P_{1,n} \neq P_{1,n-1}$, the procedure follows the second term of equation (1), looping over $l \in [1, n-2]$ attempting to find a pair $l \cdot n$ which is consistent with the value of $P_{1,n}$. Such a pair splits the sequence into two disjoint substructures on the segments $[1, l-1]$ and $[l+1, n-1]$. Both segments are pushed on the stack, and the pair $l \cdot n$ is added to $\mathcal{P}$, yielding $\mathcal{S} = ([1, l-1].[l+1, n-1]; \{l \cdot n\})$. The procedure now repeats in a similar fashion by popping the next segment from the stack. The process bottoms out as segments become too small for holding a base pair. Such segments are popped from the stack without causing other segments to be pushed. Finally, when the stack is empty, a complete structure $\mathcal{S} = (\emptyset; \mathcal{P})$ with $P_{\max} = |\mathcal{P}|$ base pairs has been reconstructed. The algorithm is sketched in table 1.

In suboptimal folding we wish to find **all** structures which meet a given suboptimality criterion. In the maximum matching case the criterion is to have **at least** $P_{\max} - \Delta$ base pairs, with $0 \leq \Delta \leq P_{\max}$. It is here that the notion of a partial structure becomes useful. $\mathcal{S} = (\sigma; \mathcal{P})$ actually represents a **set** of structures, all of which have the base pairs $\mathcal{P}$ in common. We shall call a partial structure $\mathcal{S}_1 = (\sigma_1; \mathcal{P}_1)$ a *refinement* of a partial structure $\mathcal{S}_2 = (\sigma_2; \mathcal{P}_2)$, written $\mathcal{S}_1 < \mathcal{S}_2$, if $\mathcal{P}_2 \subseteq \mathcal{P}_1$ and $\forall\ [a, b] \in \sigma_1\ \exists\ [c, d] \in \sigma_2$, such that $[a, b] \subseteq [c, d]$, with strict inequality holding at least once (otherwise $\mathcal{S}_1 = \mathcal{S}_2$).

As before, backtracking consists in the iterated refinement of the set of all structures $\mathcal{S} = ([1, n]; \emptyset)$. The difference now is that at each stage **all** refinements $\mathcal{S}'$ are kept representing (complete) structures with at least $P_{\max} - \Delta$ base pairs. To decide whether an $\mathcal{S}'$ qualifies, consider a refinement generated from $\mathcal{S} = ([i, j].\sigma; \mathcal{P})$ by splitting segment $[i, j]$ with a base pair $l \cdot j$ for some l, $i \leq l \leq j-2$. This yields the partial structure $\mathcal{S}' = ([i, l-1].[l+1, j-1].\sigma; \mathcal{P} \cup \{l \cdot j\})$. The maximum number $P_{\mathcal{S}'}$ of base pairs that any structure in the set represented by $\mathcal{S}'$ can have is given by

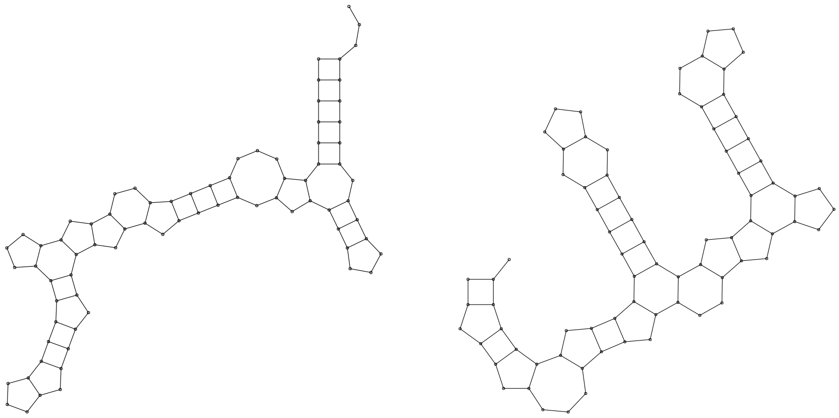
$$P_{\mathcal{S}'} = |\mathcal{P}| + 1 + P_{i,l-1} + P_{l+1,j-1} + \sum_{[a,b] \in \sigma} P_{a,b}, \tag{2}$$

where the entries of the P array have been computed in a previous optimization pass. If $P_{\mathcal{S}'}$ is less than the required minimum, $P_{\max} - \Delta$, the set $\mathcal{S}'$ can safely be pruned from further consideration. If, on the other hand,

$$P_{\mathcal{S}'} \geq P_{\max} - \Delta, \tag{3}$$

the partial structure $\mathcal{S}'$ is kept for further refinements.

Previously we backtracked (1) by refining a particular partial structure all the way down to a single complete structure. Now, in contrast, we push each refinement satisfying the suboptimality criterion (3) on a stack R of partial structures for further iterative refinement. The algorithm is summarized in table 2.



```
GGUGGCUANAGCUCAGNNGGNAGAGCCCUGGAUUNUGNUUCCAGUUNUCGUGGGNUCGAAUCCCAUUAGCCACCCCA
(((((((...(((((..((.((...)((((((...).))).)))).)).)))).).)(((...))).))))))...
((.((.(...(((.((..((.((...)((((((...)..))))))).))(((((.(...).)))))))).))))))).
```

Figure 2: **Two solutions maximizing base pairing in *E. coli* tRNA$^{\text{his}}$ (RH1660).** Certain modified bases in the sequence were replaced by a non-bonding nucleotide N, see Appendix C.

When choosing $\Delta = 0$, the algorithm produces all degenerate optimal solutions to the maximum matching problem. If $\Delta = P_{\max}$, the algorithm degenerates to a systematic construction of all admissible structures on the given sequence. The latter can take a long time, since the number of structures scales exponentially with the sequence length. We performed sanity checks of our implementation by comparing the number of admissible structures generated by it with outputs from independent structure counting procedures based on the partition function algorithm [11, 12] and a density of states algorithm [13].

For example, the maximum matching solution for the *E.coli* tRNA$^{\text{his}}$ sequence RH1660 has 26 base pairs (we require that a hairpin turn must have

```
S = {σ = {[1,n]}; P = ∅},  R = ∅
S ⇒ R
repeat:
if (R = ∅)
      done
(σ ; P) ⇐ R
if (σ = ∅)
      output suboptimal solution (∅; P) and repeat
[a,b] ⇐ σ
let S' = ([a,b−1].σ ; P)
if (P_S' ≥ P_max − Δ)
      S' ⇒ R
for each l·b with l ∈ [a,b−2]
 {
      let S' = ([a,l−1].[l+1,b−1].σ ; P ∪ {l·b})
      if (P_S' ≥ P_max − Δ)
            S' ⇒ R
 }
if ( nothing has been pushed on R since the last repeat )
      (σ ; P) ⇒ R
repeat
```

Table 2: **Pseudocode for generating suboptimal solutions to the maximum matching problem.** The meaning of $\Leftarrow$ and $\Rightarrow$ is explained in table 1. Recall that $[a,b] \Leftarrow \sigma$ deletes $[a,b]$ from σ. It is understood that in the case of illegal intervals ($[a,b]$ with $a > b$) all statements referring to that interval are skipped.

at least 3 unpaired positions). Choosing $\Delta = 0$ we find $149,126$ structures, all having the maximum of 26 base pairs. Two instances are shown in Figure 2. To find all "ground states" required 125 seconds CPU time on a SUN Ultra 2 (256 Mb memory) with our prototype implementation not tuned for efficiency. There are $9,889,659$ solutions with 25 or more base pairs (2.2 hours), and $318,369,772$ structures with 24 or more base pairs (68.6 hours).

SUBOPTIMAL FREE ENERGY FOLDING

From the thermodynamic point of view the building blocks of secondary structures are loops (Figure 3) - stacked base pairs, internal loops, bulges, and multiloops (i.e., structural elements delimited by more than one base pair) - rather than individual base pairs. As a consequence, the minimum free energy folding algorithm requires a number of distinct arrays (see below). This complicates the backtrack procedure.

Like in the previous section, the suboptimal free energy backtrack must refine partial structures by exactly reversing the optimization procedure used to systematically generate structures from smaller segments. If we were not to proceed in this way, the energy arrays filled during the optimization pass could not be used in the pruning criterion. A problem arises, however, when reversal of the optimization procedure yields more than one way of generating the *same* structure. It is in particular the construction of multiloops which needs attention in this regard. Reversing the usual Zuker-Stiegler procedure [4] yields vast amounts of structure repetitions (with the same energy) due to the non-uniqueness of their multiloop decomposition. This is obviously irrelevant when tracing back for the optimal structure, but of little use in the systematic generation of suboptimal structures. Our solution to this consists in modifying the Zuker-Stiegler procedure by decomposing multiloops in a unique way (see Appendix A).

A further problem arises from energy contributions due to so-called dangling ends. Unpaired bases adjacent to a helix may lower the energy of a structure by stacking onto their neighboring base pairs. These contributions are taken into account for external bases (a base not enclosed in any loop, see Figure 1) adjacent to the $5'$ and $3'$ end of a helical region. The same holds for unpaired nucleotides inside a multiloop which are adjacent to helical regions (Figure 3). Normally, a base may not simultaneously participate in both interactions

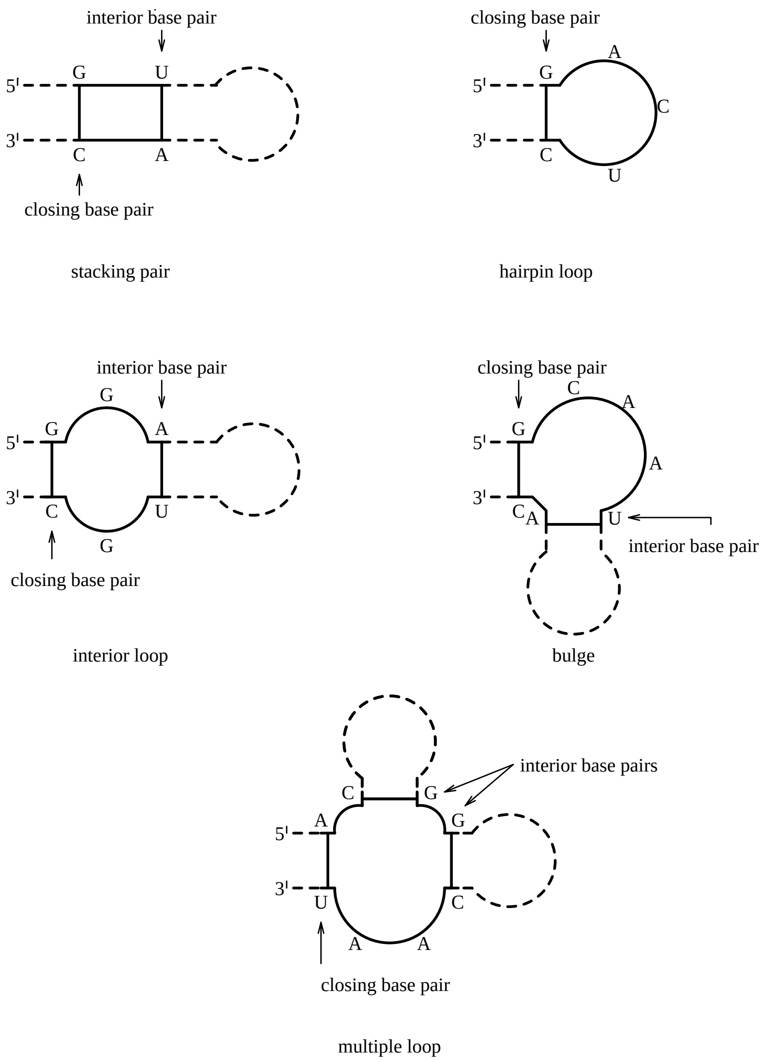


Figure 3: **Secondary structure elements.**

a $5'$ dangling end with one helix **and** a $3'$ dangling end with another. The problem for the suboptimal backtrack is that, when decomposing a multi-loop, we do not know yet whether a base adjacent to a helix is available for a dangling end interaction, that is, whether that base is unpaired and not already involved in another dangling end interaction. We handle this situation simply by always adding a dangling end contribution without checking whether the base involved qualifies. This leads to additional dangling end contributions for helices directly adjacent to one another. Incidentally, such helices often do engage in stabilizing interactions through co-axial stacking [14]. Our treatment of dangling ends, thus, can actually improve predictions in some cases. Alternatively, the contributions from dangling ends can be switched off alltogether. They can, however, be substantial. Compilations of energy parameters used in our implementation are given in [15, 16, 17, 18].

With the algorithm described in Appendix A the Waterman-Byers scheme can be used to find all suboptimal structures within a given energy range above the minimum free energy $E_{\min}$. Exactly like in the maximum matching case, we proceed by refining partial structures, and checking whether the refinements survive a pruning criterion analogous to equations (2,3). The technical definition of a partial structure needs two slight amendments. First, in addition to the set of base pairs $\mathcal{P}$ and the stack of segments σ, we keep track of the total free energy, $E_{L_\mathcal{S}}$, of all loops $L_\mathcal{S}$ which constitute a given partial structure $\mathcal{S}$. Second, each segment on the stack σ requires an additional label, indicating in which one of the arrays F^5, C, F^M, and F^{M1} (see Appendix A) the best energy attainable on that segment should be looked up. These labels are also needed in switching the backtrack between the appropriate arrays. The labels are assigned according to how a segment is generated through refinement from another segment (see Appendix B).

Let us denote the best possible energy attainable on segment $[i,j]_E$ with $E_{i,j}$, where $E \in \{F^5, C, F^M, F^{M1}\}$. Suppose now that we are refining a partial structure $\mathcal{S} = ([i,j]_E.\sigma; \mathcal{P}; E_{L_\mathcal{S}})$ by reversing the minimization procedure outlined in table 4 of Appendix A. Depending on the equation used, this will yield subintervals of $[i,j]_E$ whose total best possible energy we denote here with $\Box$ to avoid distracting details. All (complete) structures represented by the attempted refinement of $\mathcal{S}$ have an energy not lower than $E_\mathcal{S}$, where

$$E_\mathcal{S} = \Box + E_{L_\mathcal{S}} + \sum_{[k,l]_E \in \sigma} E_{k,l}. \tag{4}$$

sequence length	range of energy/kT						
	5	10	12	15	17	20	
25	17	187	441	1299	2569	6048	structures
	0	0	0	0	0	0	CPU secs
50	9	108	254	900	2178	6477	structures
	0	0	0	0	0	2	CPU secs
75	86	1664	5056	24299	67601	295722	structures
	0	1	2	10	34	201	CPU secs
100	121	4439	16567	103935	341054	1864633	structures
	1	6	10	54	169	1815	CPU secs

Table 3: **CPU time of the suboptimal folding algorithm.** Calculations were made on a DEC AlphaServer 2100 5/375.

In analogy to equation (3), we will accept any refinement for which

$$E_{\mathcal{S}} \leq E_{\min} + \delta, \tag{5}$$

with some desired $\delta > 0$. The strategy for tracing back all suboptimal structures in the energy range between $E_{\min}$ and $E_{\min} + \delta$ is detailed in Appendix B. The logic is the same as in the maximum matching case, but the details are more sophisticated.

Again, a choice of $\delta = 0$ yields a conventional backtrack with the added benefit of finding all degenerate "ground structures", should there be more than one. Choosing δ large enough, say $\delta = \infty$, makes the algorithm degenerate again into a structure counting procedure. This is quite handy for a basic soundness check by comparing whether its output coincides with that of the maximum matching algorithm with $\Delta = P_{\max}$.

Performance considerations

The time to compute all structures with energy in the interval between $E_{\min}$ and $E_{\min} + \delta$ trivially depends on how many structures this interval contains. The relevant point for practical purposes comes from full density of states calculations (providing energy levels, but no structures) [13] which suggest that the number of states is rather modest around $E_{\min}$, and typically blows up only at energies substantially higher than $E_{\min}$ (see, for example, table 3).

This is a welcome contrast to the maximum matching case. Thus, as long as δ is small (say within a few multiples of kT), our procedure is extremely fast.

Table 3 summarizes CPU requirements of the algorithm for 4 test sequences of lengths 25, 50, 75 and 100 at various energy intervals above E_{min} in multiples of kT. The data show exponential behavior in this energy range with regard to number of structures and CPU time. Memory demands remain modest even for large sequences and δ. Normally a sorted list of the calculated structures is desirable, in which case the sorting becomes the dominat time and memory factor.

THERMODYNAMIC STABILITY OF tRNA CLOVERLEAF STRUCTURE

An example from yeast

Using the suboptimal folding algorithm just developed, the 50 energetically lowest structures of the yeast $\mathrm{tRNA}^{\mathrm{phe}}$ (RF6280) were generated. Figure 4 shows a clustering of that set of structures based on Ward's variance criterion [19]. The procedure starts out with each structure being a cluster. At each iteration two clusters are merged into a larger one so as to minimize the associated increase in variance. Computing the variance requires a notion of distance between two structures. Here we take distance to be the total number of base pairs which **both** structures do not have in common, that is, the symmetric difference between their sets of base pairs. For example, "`((((....))))`" and "`.((((...))))`" have a distance of 8 (each structure has 4 base pairs which the other hasn't), while "`((((....))))`" and "`.(((....))).`" have a distance of 1 (the former structure has 1 base pair which the latter lacks, but the latter has no base pair missing in the former).

The input sequence for Figure 4 was obtained from the original RF6280 by replacing all modified bases with their unmodified analogues. We shall refer to sequences of this kind as "unmodified sequences". A second "modified" input sequence was obtained by translating a subset of modified bases into a non-bonding nucleotide following [20, 21, 22] (see Appendix C). The structures of the unmodified sequence which coincide with the 12 best structures of the modified one are highlighted in bold face (Figure 4).

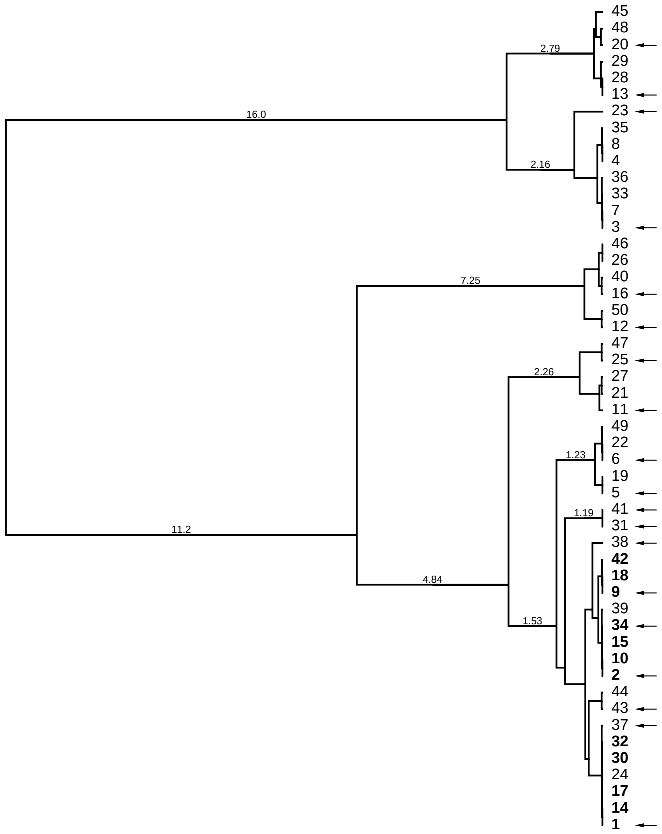


Figure 4: **Similarity clustering of suboptimal structures of yeast tRNA$^{\mathrm{phe}}$.** The 50 energetically lowest structures, ranging from -19.26 kcal/mol to -17.28 kcal/mol, of yeast tRNA$^{\mathrm{phe}}$ (EMBL accession RF6280) are clustered using Ward's variance method [19] with the symmetric difference distance as a metric on the structures. Numbers at the leaves of the tree indicate the energy rank of a suboptimal structure (mfe structure is #1). The structures belong to the sequence obtained from RF6280 by translating modified bases into their corresponding unmodified analogues. Numbers in bold face flag structures which coincide with the 12 lowest suboptimal structures of the sequence obtained from RF6280 by replacing certain modified bases by a non-bonding nucleotide. Arrows indicate structures satisfying the definition of Zuker's suboptimal folding scheme [6].

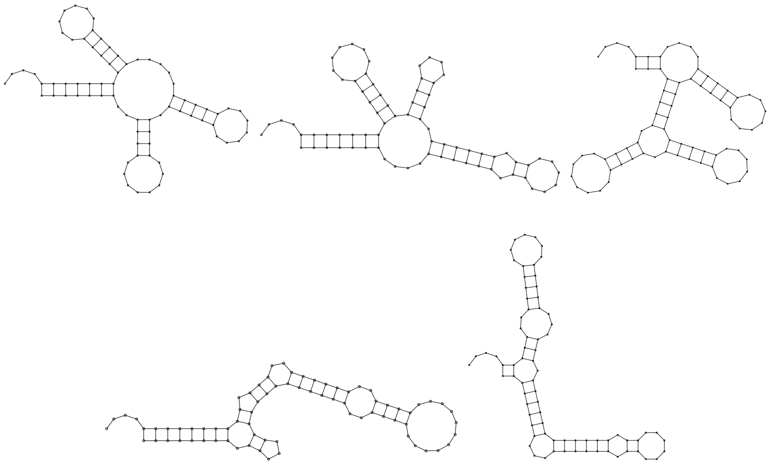
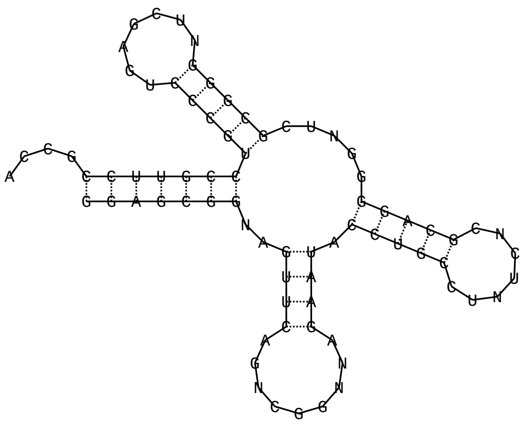


Figure 5: **Structural diversity near the mfe of yeast tRNA$^{\mathbf{phe}}$.** The structures shown correspond to the cluster nucleators #1,#3,#11,#12,#13 (from left to right, top to bottom) in Figure 4.

The first point made by this example is the existence of structures in the neighborhood of the mfe structure which differ substantially from it. The data of Figure 4 show that the low energy region of the unmodified tRNA^{phe} comprises at least two major classes of structures. In particular, the mfe structure (-19.26 kcal/mol) is in one class, while a structure as close to it as #3 (-18.83 kcal/mol) belongs to a different class. These classes are split into further clusters, and Figure 5 gives an indication of their structural diversity. This is a static picture, and nothing is said about the barrier between #1 and #3. By systematically generating the complete configuration space around the mfe, our procedure can assist in obtaining either the barrier itself or a lower bound to it. However, we shall not be concerned with kinetics in this communication.

Both modified and unmodified sequences fold into the same mfe structure, and the 12 structures with lowest energy of the modified sequence are among the 50 lowest structures of the unmodified variant. However, all 12 structures of the modified sequence group into the same cluster. This raises the issue about the effect of tRNA base modification on the density and diversity of states around the mfe. In cases where the unmodified sequence folds into the correct cloverleaf structure, modifications that prevent base pairing do not alter the mfe structure. They seem, however, to constrain structures at low energies to be similar to the ground state. This suggests that base modification improves the "definition" of the mfe structure. We shall return to this point in greater detail.



```
RI1660 AGGCUUGUAGCUCAGGDGGDDAGAGCGCACCCCUGAU6AGGGUGAG7XCGGUGGTPCAAGUCCACPCAGGCCUACCA
RI1661 AGGCUUGUAGCUCAGGUGGDDAGAGCGCACCCCUGAU6AGGGUGAG7XCGGUGGTPCAAGUCCACPCAGGCCUACCA
RR1661 GCAUCCG4AGCUCAGCDGGADAGAGUACUCGGCUICG/ACCGAGCG7XCGGAGGTPCGAAUCCUCCCGGAUGCACCA
RV1660 GCGUCCG4AGCUCAGDDGGDDAGAGCACCACCUUGACAUGGUGGGG7XCGGUGGTPCGAGUCCACUCGGACGCACCA
RV1661 GCGUUCA4AGCUCAGDDGGDDAGAGCACCACCUUGACAUGGUGGGG7XCGUUGGTPCGAGUCCAAUUGAACGCACCA
RD1660 GGAGCGG4AGUUCAGDCGGDDAGAAUACCUGCCUQUC/CGCAGGGG7UCGCGGGTPCGAGUCCCGPCCGUUCCGCCA
```

Figure 6: **tRNA secondary structure shared by six sequences from *E.coli*.** The figure shows the secondary structure (obtained with the thermodynamic folding algorithm) shared by the six listed tRNA sequences. (The second letter in the accession number identifies the amino acid.) The structure has the tRNA^{asp} sequence (RD1660) superimposed, indicating the positions at which a non-bonding base **N** was placed. The shaded areas in the list of sequences indicate the positions at which **N**s were placed in inverse folded sequences of the modified sample.

Figure 4 also indicates that Zuker's subset of suboptimal structures [6] (those that are optimal with respect to the choice of a base pair) does indeed constitute a representative sample of the structural variability in the vicinity of the mfe (shaded areas in Figure 4).

Diversity of states in modified and unmodified artificial tRNAs

The previous example suggests a role for modified bases in altering the structural states in the vicinity of the mfe. Yet, conclusions which rest on the details (in particular the ordering) of these states for a single sequence remain susceptible to the same imprecisions in the available energy parameters as pure mfe-folding. One way around this problem is to turn away from the structure prediction and analysis of a single sequence to a statistical approach [23] in which we identify and compare robust properties of specific (natural or artificial) sequence ensembles. This approach can be expected to yield conclusions that are robust to variations in the energy parameters.

Using an inverse folding procedure [24, 25] we generated a pool of 2000 sequences whose mfe structure coincides with that of the six natural sequences listed in Figure 6. This constitutes an ensemble of unmodified sequences, or "unmodified sample" for short. Similarly, we generated a pool of 2000 sequences with a non-bonding nucleotide at every position indicated in Figure 6. These sequences were chosen so as to have the same mfe structure as the unmodified pool. We shall refer to this ensemble as the "modified sample".

For each sequence in both samples we computed the energy gap between the mfe structure and the second best structure. The distribution of these first gap energies is shown in Figure 7. An immediate observation is that natural tRNA sequences have large first gap energies, located far out in the tail of the distribution. A more subtle feature, however, is that the modified sample exhibits a set of spikes rising distinctively above a generally flatter background as compared to the unmodified sample. An analysis of the structures associated with the gap energies at these spikes reveals that the extent to which all major spikes rise above the background is due precisely to those structures resulting from the ground state by removing one base pair at either end of a helical region. For example, 74%, 70%, 80%, 92%, and 88% of the structures at gap energies 0.09, 0.17, 0.29, 0.69, and 0.97 kcal/mol above the mfe, respectively, lack either one base pair at the acceptor end of the multiloop or at the loop end of one of the hairpin turns. This "quantized" superstructure in the gap distribution of the modified sample shows that non-bonding bases constrain the second best structure to be as similar as possible to the ground state. This is in marked contrast to the unmodified sample, where larger refolds at the first energy level are considerably more likely. Consider that the energy difference between the ground structure and

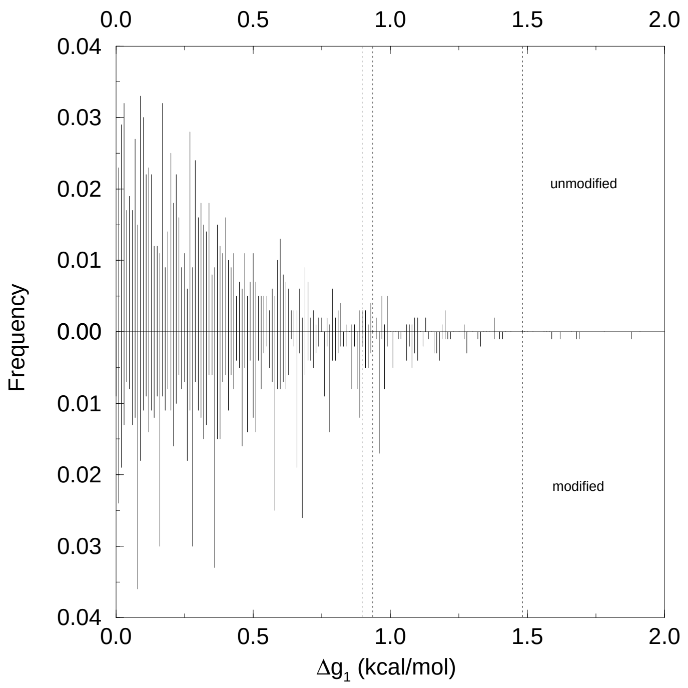


Figure 7: **Distribution of first gap energies.** The upper and lower half show the distribution of gap energies in the unmodified and modified sample, respectively. The dotted vertical lines indicate the gap energies of the six natural sequences in Figure 6

the second best structure cannot be larger than the largest stacking energy ($-\Delta G_{\mathrm{GC\cdot CG}}$, which is about 3 kcal/mol at 37° C). To see this, assume that S is the second best structure. If S were to differ more than the stated amount from the mfe structure, we could construct a better structure by simply removing a base pair at either end of some stack in the mfe structure, thus contradicting the assumption that S is the second best structure. Hence, the next structure above the ground state will be a similar structure with just a base pair removed from a helix end only if there exists no refolded configuration with a lower energy. Figure 7 shows that properly placed non-bonding bases make the latter possibility distinctively less likely.

Structural stability of secondary structures

In view of the previous analysis we ask whether the density of states at low energies and their associated structures can be used to quantify the degree to which an mfe structure is "well-defined". Intuitively, and from a static viewpoint, a structure is well defined if there are no "substantially different" structures in its thermodynamic neighborhood. Even in the absence of a kinetic assessment, criteria of static well-definition can be useful in identifying parts of an RNA structure with biological significance.

Extant measures of well-definition use McCaskill's partition function algorithm [11]. For example, one may quantify the most likely state - paired or unpaired - of a position k in a sequence by the probability of the most probable base pair involving k, or the probability that k is unpaired, whichever is larger [26]. These base pair probabilities are obtained from the partition function Z [11], which can also be approximated with the suboptimal folding procedure by summing over the density of states at low energies [27].

At the other extreme one might consider a simple global measure as given by the fraction of the mfe structure in the Boltzmann ensemble:

$$f_{\mathrm{mfe}} = \frac{e^{-\Delta G_{\mathrm{mfe}}/kT}}{Z} = \frac{1}{1 + \sum_i e^{-\Delta g_i/kT}}, \tag{6}$$

where Δg_i is the ith gap energy $\Delta G_i - \Delta G_{\mathrm{mfe}}$. This can also be expressed as $kT \ln f_{\mathrm{mfe}} = F - \Delta G_{\mathrm{mfe}}$, where F is the free energy of the Boltzmann ensemble. Another such measure is the mean gap energy $< \Delta g >$:

$$< \Delta g > = \sum_i (\Delta G_i \; - \; \Delta G_{\mathrm{mfe}}) \, \frac{e^{-\Delta G_i/kT}}{Z}. \tag{7}$$

Figure 8 shows the distribution of f_{mfe} in the modified and unmodified samples, together with the values for the natural sequences of Figure 6. Again, the latter show a remarkably high f_{mfe} as compared to both samples with the same mfe structure. The comparison between the two samples evidences the role of modified bases in shifting f_{mfe} to higher values. The bad news, however, is that a high f_{mfe} does not imply a large separation to the energetically adjacent structure, although the reverse is true (Figure 9).

Neither the quantified pairing state of an individual position, nor f_{mfe} provide sufficient information about *structural* stability. The former is too local a

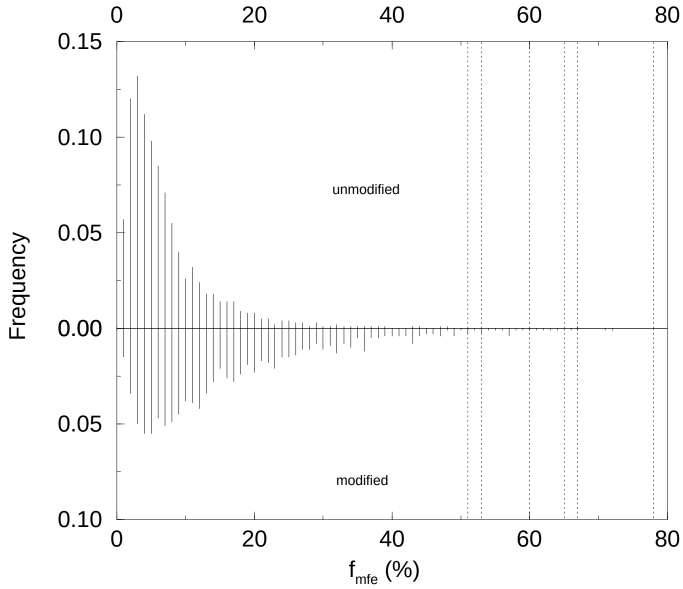


Figure 8: **Fraction of mfe structure in the Boltzmann ensemble.** The upper and lower half show the distribution of f_{mfe} in the unmodified and modified sample, respectively. The dotted vertical lines indicate the f_{mfe} of the six natural sequences of Figure 6

measure to say much about structural diversity in the vicinity of the mfe, and a low f_{mfe} can be caused by a number of similar structures that are energetically nearby the mfe. Yet, in the latter case, we would still consider the basic architecture of the mfe to be well-defined.

A simple measure for the structural diversity present in the secondary structure configuration space of a sequence is the Boltzmann weighted sum over the structure distances between the ith configuration and the ground state. As a structure distance we use the so-called base pair distance, defined as follows: each position in structure A that is not paired to the same position as in structure B increases the distance by one count. In this metric one-strand shifts of helical regions give large distances. For example, "`((((....))))`" and "`.((((...))))`" have a base pair distance of 9 (all 8 paired positions differ), while "`((((....))))`" and "`.(((....))).`" have a base pair distance of 2. (Base pair distance is similar but not identical to the symmetric

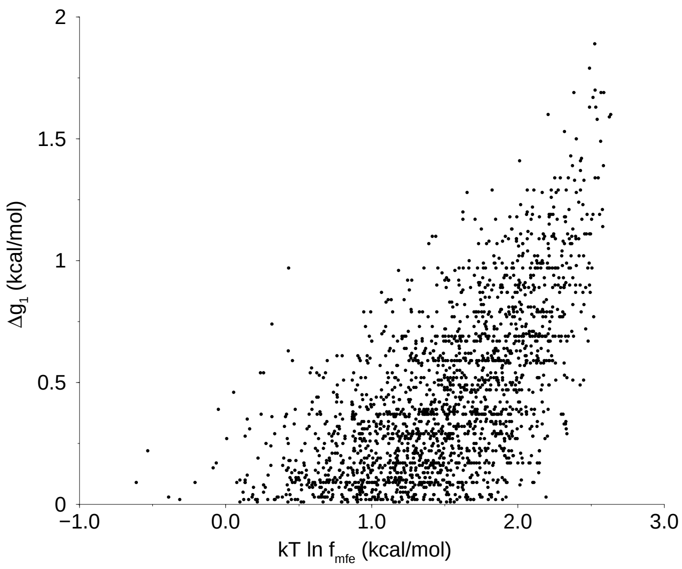


Figure 9: **Relation between f_{mfe} and the first gap energy.** The figure shows a scatter plot where the first gap energy of each sequence in the modified sample is plotted against its f_{mfe}. Recall that an upper bound for Δg_1 is about 3 kcal/mol. The figure shows that a high Δg_1 implies a high mfe fraction in the Boltzmann ensemble, but that the reverse is not true. A similar picture holds for the unmodified ensemble.

difference distance used earlier in this section.)

$$< d_{\mathrm{bp}} > = \sum_i d_{\mathrm{bp}}(0,i)\, \frac{e^{-\Delta G_i/kT}}{Z}, \tag{8}$$

where $d_{\mathrm{bp}}(0,i)$ denotes the base pair distance between the mfe structure (0) and the ith structure above it.

In Figure 10 we plot for three classes of tRNA sequences the mean gap energy against the mean base pair distance, which we approximated by considering all structures within $10kT$ of the mfe. The three classes were derived from Steegborn's compilation [28] of *E. coli* tRNA sequences. The first class (filled circles in Figure 10) consisted of the natural tRNA sequences whose modified bases were replaced by non-bonding bases in their original positions (accord-

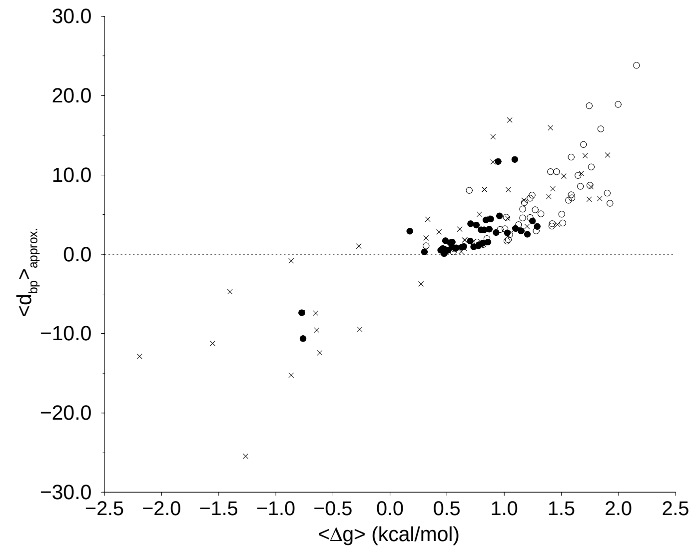


Figure 10: **Structural diversity versus mean gap energy.** The figure plots a $10KT$ approximation to the mean structure distance, as defined in equation (8), and the mean gap energy, as defined in equation (7), for *E. coli* tRNA sequences from the Steegborn compilation [28]. Filled circles: natural modified sequences. Open circles: sequences with non-bonding nucleotides at random positions but preserving the mfe structure of their natural modified counterparts. Crosses: unmodified sequences.

ing to the translations of Appendix C), the second class (open circles) had the same amount of non-bonding bases, this time in random positions, but so as to yield the same mfe structure as the originals. The third class (crosses) had the modified bases replaced by their corresponding unmodified ones. Not all of the latter had the same mfe structure as their native (i.e., modified) counterpart. (The algorithm missed the cloverleaf also for a few modified sequences.) In such cases we took the lowest lying cloverleaf structure as the reference (0). As a consequence the mean gap energy can become negative. For better readability of the plot, we assigned the mean base pair distance the same sign as the mean gap energy (but it obviously means a positive value).

Figure 10 shows that the natural modified sequences (filled circles) have by and large very small $< d_{\rm bp} >$ values indicating that most structures close

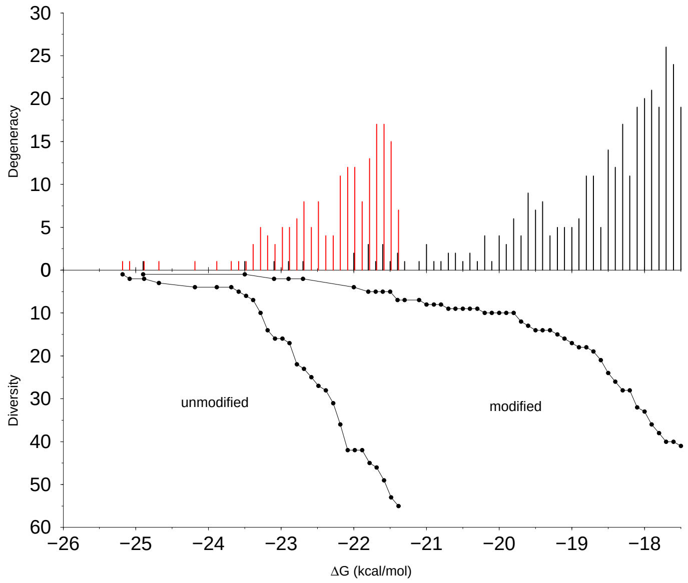


Figure 11: **Diversity of coarse grained structures.** The upper half shows the density of states for the unmodified sample (up to $15kT$) on the left, and for the modified sample (up to $30kT$) on the right. The lower half is a plot of the cumulative number of different coarse grained architectures encountered with increasing energy; left curve: unmodified sample, right curve: modified sample.

to the mfe structure are also similar to it. At the same time the mean gap energy has a wide spread. This shows that $< d_{\rm bp} >$ is a better predictor of structural stability than $< \Delta g >$. The same trend is confirmed by plots similar to Figure 10 for the modified and unmodified samples (not shown). Furthermore, sequences with non-bonding bases at random positions (open circles) have a better defined mfe structure than unmodified sequences, but not as well defined as the originals. Thus, the positioning of the non-bonding bases is important, even when it does not affect the mfe structure itself. It is tempting to interpret these data as natural tRNA sequences having their non-bonding bases positioned so as to also maximize the definition of the ground state structure.

A further assessment of structural well-definition is obtained by counting the number of different structure "architectures" as energy increases from the ground state. By "architecture" we mean a coarse-grained secondary structure obtained by disregarding the size of loops and helices [29]. Such a coarse-grained structure constitutes an equivalence class of conventional secondary structures with respect to the topological arrangement of loops and helices. The upper plot of Figure 11 shows the density of states at low energies, that is, the number of states existing at any given energy up to $15kT$ from the ground state for the unmodified *E. coli* $\mathrm{tRNA}^{\mathrm{lys}}$ sequence (RK1660), and up to $30kT$ for the modified sequence. The lower half of the plot displays the cumulative count of different coarse grained structures encountered since the mfe structure. The difference in the rates by which structural diversity increases is quite impressive, as is the difference in energy from the ground state at which diversity starts rising fast. This indicates once more that, from a thermodynamic point of view, the modified sequence is structurally much more stable. We found similar observations to hold for the other *E. coli* tRNA sequences as well.

Finally, Figure 12 shows an intriguing relationship between the thermodynamic stability of a structure and the fraction of neutral mutants accessible by a single point mutation. Neutral here means that reference sequence and the mutant have the same mfe structure. Each part of Figure 12 plots for each sequence in the modified and unmodified sample the logarithms of f_{mfe}, $< d_{\mathrm{bp}} >$, and $< \Delta g >$ against the fraction of neutral mutants of that sequence. For both samples there is a clear correlation between well-definition of the ground state and the degree to which a sequence can buffer mutations against altering that ground state. Average neutrality is higher for modified sequences, since their mfe structure is on average thermodynamically better defined than for unmodified sequences. The best predictor of mutational stability is again the mean base pair distance, while the mean gap energy is virtually insensitive. Small mean base pair distance (high thermodynamic structural stability) implies high neutrality, but the reverse, while true to some degree for unmodified sequences, does not hold for modified ones. Given that properly modified sequences have intrinsically a better defined ground state, a high degree of neutrality does no longer discriminate between different degrees of well-definition within that sample.

CONCLUDING REMARKS

Following an idea of Waterman and Byers [9] we have devised and implemented an algorithm which rigorously generates all energetically suboptimal

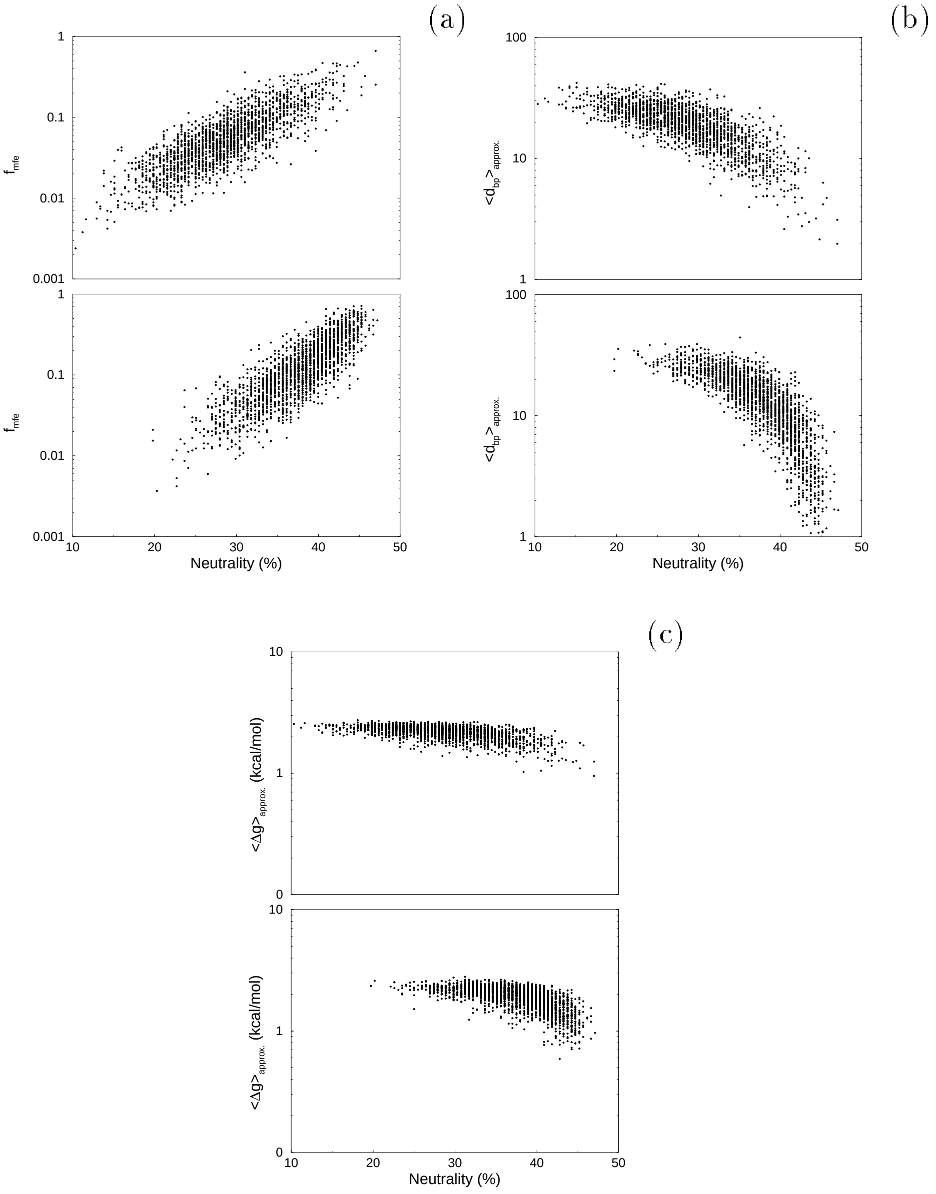


Figure 12: **Relationship between thermodynamic stability and neutrality.** Each graph (a), (b), and (c) plots one measure of thermodynamic well-definition of the ground state versus the fraction of neutral mutants accessible by one point mutation for each sequence in the unmodified sample (upper plot in each part) and the modified sample (lower plot in each part).

secondary structures of an RNA sequence within a desired energy range above the minimum free energy. The logic of the algorithm was discussed for the simple case of base pair maximization. To implement a suboptimal folding procedure based on the free energy of structures, we had to modify the Zuker-Stiegler strategy for free energy minimization. Minimization and suboptimal backtrack are detailed in the appendices.

Depending on the choice of energy range, the algorithm has two limiting behaviors. If the interval above the minimum free energy is set to zero, all degenerate ground states are obtained, while a sufficiently high energy range yields a systematic structure counting procedure. Since the density of states is relatively sparse in the 10-15kT vicinity of the minimum free energy, the algorithm is fast and practical even for long sequences. Our implementation and all algorithms used in this paper are freely available for academic research [30], and will be integrated in the next release of the Vienna RNA Package [24].

A suboptimal folding algorithm that generates rigorously all suboptimal configurations between the minimum free energy and some chosen upper limit is important for a meaningful approximation of statistical quantities. Because of this property our algorithm has the pleasent feature that energy minimization, suboptimal structures, the (truncated) density of states, the (truncated) partition function (and other statistical quantities derived from it) are unified in a single procedure and obtainable in the same optimization plus backtracking pass.

In the second part of this contribution we used our procedure to compute indicators for the thermodynamic stability of the minimum free energy structure of an RNA sequence. We defined three simple indicators capturing in different ways the degree of well-definition or well-determination of the ground state structure: (i) mean gap energy, that is, the average energy separation of configurations in the vicinity of the minimum free energy, (ii) Boltzmann weighted mean structure distance (here implemented as mean base pair distance), that is, the average distance between the minimum free energy structure and the configurations in its energy neighborhood, and (iii) topological diversity, that is, the number of different coarse grained structures in an energy interval around the ground state. These quantities were used to assess the influence of base modification on the thermodynamic robustness of the ground state structure in tRNA sequences. To this end we compared the statistics of these indicators in large samples of modified and unmodified artificial sequences whose minimum free energy structure is identical to that of naturally occurring tRNA sequences from *E. coli*. The latter were also studied individually. Base modification was considered here only in its quality of preventing particular positions in the linear sequence from contributing base pairs to the secondary structure.

Our study shows from several perspectives that base modification considerably sharpens the definition of the ground state structure by constraining energetically adjacent structures to be similar to the ground state. Base pair distance turned out to be the best indicator for how well the ground state is determined. Artificial sequences with non-bonding nucleotides at random positions, yet with the natural tRNA cloverleaf pattern as ground state, determine the cloverleaf better than unmodified sequences, but not as well as natural sequences with the same secondary structure. This indicates that certain positions when locked into a non-bonding state are more effective than others in sharpening the thermodynamic definition of the minimum free energy structure.

There is a noteworthy correlation between the thermodynamic stability of the minimum free energy structure of a given sequence and its capacity to buffer mutations. The better the ground state is defined, the more one-error mutants preserve the minimum free energy structure. This may have evolutionary consequences at the molecular level. If well-definition of a secondary structure is important for biological function, then evolving a sequence which improves the thermodynamic definition of that structure has as a likely side effect an increased stability towards point mutations, that is, neutrality.

The importance of a rigorous suboptimal folding algorithm rests not only with computing criteria for discerning biologically relevant structures held under selection pressure, or for detecting relevant alternative states to the ground state. A key issue will be to unravel the kinetic aspects of RNA folding, and to understand what makes a sequence fold well. By providing access to the complete configuration space at low energies, we expect a rigorous suboptimal folding algorithm to be a valuable tool towards that goal.

ACKNOWLEDGEMENTS

Several drawings were kindly provided by Jan Cupal. This work was supported by the Austrian Fond zur Förderung der wissenschaftlichen Forschung, project No. 11065-CHE, and by the integrative core research of the Santa Fe Institute.

APPENDICES

A OPTIMIZATION WITH UNIQUE MULITLOOP DECOMPOSITION

In this appendix we explain the modified optimization procedure on which we base the suboptimal backtrack detailed in Appendix B.

We recall here for later reference the usual treatment for multiloop energies $\mathcal{M}$ [5],

$$\mathcal{M} = \mathcal{M}_C + \mathcal{M}_I \cdot \#\text{ interior pairs} + \mathcal{M}_B \cdot \#\text{ unpaired bases}, \tag{9}$$

where $\mathcal{M}_C$ denotes the stabilizing energy deriving from the multiloop closing pair, $\mathcal{M}_I$ denotes the stabilizing energy for each base pair interior to the multiloop, and $\mathcal{M}_B$ the destabilizing energy for each unpaired base in the loop [17].

As in the maximum matching case, energy arrays are filled in a recursive fashion. Let $C_{i,j}$ be the minimum free energy on the segment $[i, j]$, provided that i and j pair with one another. As is well known [5], by virtue of the additivity of loop energies, the best energy attainable on the segment $[i, j]$, with $i \cdot j$, is given by the energy of the particular loop L closed off by $i \cdot j$ plus the energy of any substructures ending with a base pair $p \cdot q$ in that loop.

$$C_{i,j} = \min_{\substack{\text{loops } L \\ \text{closed by } i,j}} \Big\{E(L) + \sum_{\substack{\text{interior pairs} \\ p\cdot q\,\in\, L}} C_{p,q}\Big\} \tag{10}$$

with $C_{i,i} = \infty$. The minimization in equation (10) runs over three major classes of structures, consisting of various loop types closed off by $i \cdot j$ (see Figure 13 for a schematic representation).

$$\begin{aligned} C_{i,j} = \quad \min\Big\{ & \\ & \mathcal{H}(i,j), \\ & \min_{\substack{p\in[i+1,j-m-2] \\ q\in[p+m+1,j-1]}} \big\{C_{p,q} + \mathcal{I}(i,j,p,q)\big\}, \\ & \min_{k\in[i+1,j-m-2]} \big\{F^{M}_{i+1,k-1} + F^{M1}_{k,j-1} + d^5_{i,j,j-1} + d^3_{i,j,i+1} + \mathcal{M}_C\big\} \\ \Big\} & \end{aligned} \tag{11}$$

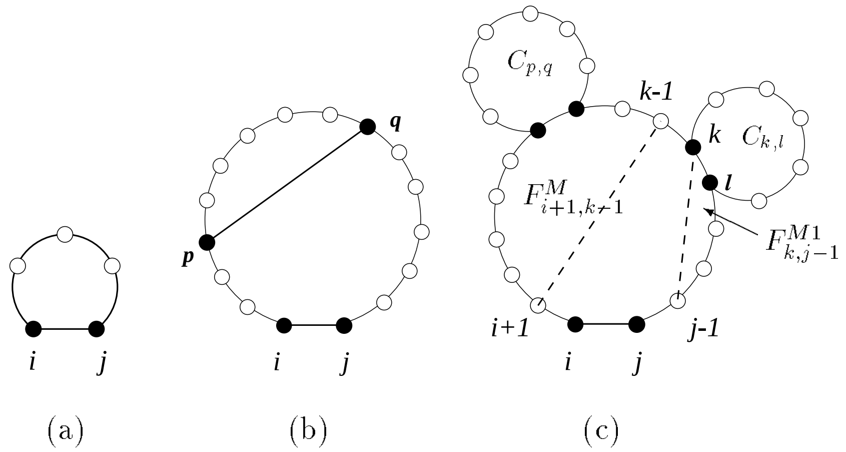


Figure 13: **Schematic representation of the terms in equation (11).** First term (a): base pair $i \cdot j$ closes a hairpin loop of a certain size. The minimal loop size is 3. Second term (b): base pair $i \cdot j$ closes an interior loop, whose inner base pair is $p \cdot q$. All possible pairs $p \cdot q$ must be considered. Third term (c): base pair $i \cdot j$ closes a multiloop with a certain number of interior base pairs (filled circles, such as $k \cdot l$). Multiloops are divided recursively into substructures, containing the last stem at the $3'$ end (energy $F^{M1}_{k,j-1}$) and the remaining $5'$ structure (energy $F^M_{i+1,k-1}$). The remaining structure is again split in the same way, see equation (13). Dangling end contributions are not shown.

The first term, $\mathcal{H}(i,j)$, denotes the tabulated free energy of a hairpin loop closed by $i \cdot j$. The second term considers all cases where $i \cdot j$ closes an interior loop (or a bulge) whose interior delimiting base pair is $p \cdot q$. The loop has a tabulated energy $\mathcal{I}(i,j,p,q)$, the structure "behind" $p \cdot q$ has energy $C_{p,q}$, and the minimum is taken over all admissible pairs $p \cdot q$. The third term refers to multiloop structures closed by $i \cdot j$. A multiloop is constructed from two pieces with energy $F^M_{i+1,k-1}$ and $F^{M1}_{k,j-1}$ (to be explained shortly; see also Figure 13), and the multiloop closing pair $i \cdot j$ with energy $\mathcal{M}_C$ (see equation (9)). We also take into account the stabilizing energy from dangling ends on the $5'$ and the $3'$ side of the pair $i \cdot j$. $d^5_{i,j,j-1}$ denotes the energy contribution of the base at position $j-1$ stacking from the $5'$ direction onto the pair $i \cdot j$. Similarly for d^3.

As indicated in Figure 13(c), $F^{M1}_{i,j}$ denotes the minimum free energy of the last stem (closed, say, by $i \cdot l$) towards the $3'$ end of the multiloop being considered, including an arbitrary number of unpaired bases at the $3'$ end. Its energy is the sum of the energy $C_{i,l}$ of the structure closed by $i \cdot l$, the energy $\mathcal{M}_B(j-l)$ of $j-l$ unpaired multiloop bases, the multiloop energy contribution $\mathcal{M}_I$ deriving from the interior pair $i \cdot l$, and the dangling ends:

$$F_{i,j}^{M1} = \min_{l \in [i+m+1,j]} \left\{ C_{i,l} + \mathcal{M}_B(j-l) + d^5_{i,l,i-1} + d^3_{i,l,l+1} + \mathcal{M}_I \right\} \tag{12}$$

The remaining $5'$ piece of the multiloop structure is further split recursively into a $3'$ stem plus a remaining $5'$ section. The recursion bottoms out when no $3'$ stem is possible. In terms of energies:

$$F_{i,j}^{M} = \min \Big\{ \min_{k \in [i+m+1,j-m-1]} \{ F^M_{i,k-1} + F^{M1}_{k,j} \}, \tag{13}$$

$$\min_{k \in [i,j-m-1]} \{ F^{M1}_{k,j} + \mathcal{M}_{\mathcal{B}}(k-i) \} \Big\}. \tag{14}$$

This procedure ensures that there is only one decomposition of a multiloop into substructures, thus enabling a meaningful suboptimal backtrack.

Finally, all we need is the best free energy on a segment $[1,j]$, denoted by F^5_j, irrespective of whether position j is paired. F^5_j is constructed recursively, as illustrated in Figure 14,

$$F^5_j = \min \Big\{ F^5_{j-1}, \min_{l \in [1,j-m-1]} \{ F^5_{l-1} + C_{l,j} + d^5_{l,j,l-1} + d^3_{l,j,j+1} \} \Big\}. \tag{15}$$

The first term represents the case where j is left unpaired. The second term considers all possible positions l that might be paired to j. The free energy $E_{\min}$ of the best structure on the entire sequence is then given by $E_{\min} = F^5_n$.

Table 4 summarizes the algorithm for computing the minimum free energy on a given RNA sequence. It has complexity $O(n^3)$, and its implementation is very fast. A structure corresponding to the minimum free energy is again obtained by backtracking through the various arrays.

In Appendix B we detail the trace back yielding all suboptimal structures with energies between $E_{\min}$ and $E_{\min} + \delta$, with $\delta > 0$ chosen by the user.

B SUBOPTIMAL BACKTRACK

We label segments $[i,j]$ with subscripts F, C, M, and $M1$, referring to the arrays F^5, C, F^M, and F^{M1}, respectively. As usual, the backtrack starts

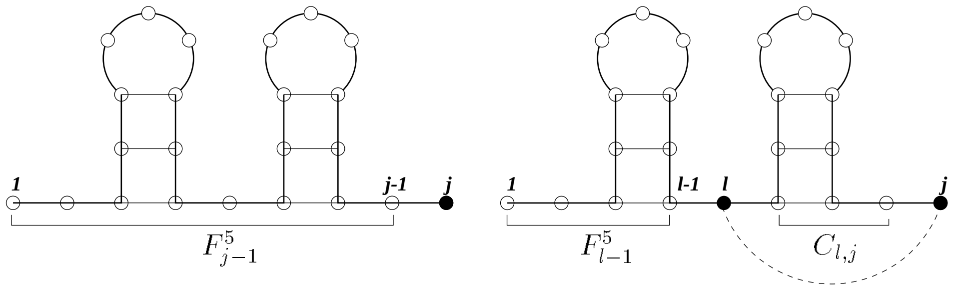


Figure 14: **Schematic representation of equation (15).** The left scheme and right scheme represent the first and second term, respectively, in the minimization of equation (15). Dangling end contributions are not indicated.

$$C_{i,j} = \min\Big\{\mathcal{H}(i,j), \min_{\substack{k\in[i+1,j-m-2]\\ l\in[k+m+1,j-1]}} \{F^B_{k,l} + \mathcal{I}(i,j,k,l)\}, \min_{k\in[i+1,j-m-2]} \{F^M_{i+1,k-1} + F^{M1}_{k,j-1} + \mathcal{M}_C\}\Big\}$$

$$F^{M1}_{i,j} = \min_{l\in[i+m+1,j]} \Big\{ C_{i,l} + d^5_{i,l,i-1} + d^3_{i,l,l+1} + \mathcal{M}_B(j-l) + \mathcal{M}_\mathcal{I} \Big\}$$

$$F^M_{i,j} = \min\Big\{ \min_{k\in[i+m+1,j-m-1]} \{F^M_{i,k-1} + F^{M1}_{k,j}\}, \min_{k\in[i,j-m-1]} \{F^{M1}_{k,j} + \mathcal{M}_B(k-i)\}\Big\}$$

$$F^5_j = \min_{l\in[1,j-m-1]} \Big\{ F^5_{j-1}, F^5_{l-1} + C_{l,j} + d^5_{l,j,l-1} + d^3_{l,j,j+1} \Big\}$$

Table 4: **Recursive calculation of the minimum free energy.** Calligraphic symbols denote tabulated energy parameters for different loop types. Hairpin loops: $\mathcal{H}(i,j)$, interior loops, bulges, and stacks: $\mathcal{I}(i,j,k,l)$; the multi-loop energy is modeled by the linear ansatz of equation (9). The particular recursion on the multiloop arrays F^M and F^{M1} yields a unique decomposition. The overall calculation proceeds from smaller segments to larger ones. The minimum free energy on the segment $[1,j]$ is stored in F^5_j. Upon completion the minimum free energy is in F^5_n.

with $\mathcal{S} = ([1,n]_F; \emptyset; 0)$. We outline the procedure involved in refining the partial structure $\mathcal{S} = ([i,j]_E.\sigma; \mathcal{P}; E_{L_S})$ which has just been popped from the partial structure stack R. The segment $[i,j]_E$ is popped from the partial structure's segment stack, and refined according to the marker E.

▶ case $E = F$ (backtrack in F^5)

i and j are external bases, and the possible refinements follow equation (15). Leaving the $3'$ end unpaired, leads to the acceptance condition

$$F^5_{j-1} + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \leq E_{\min} + \delta. \tag{16}$$

If (16) is fulfilled, we push the new partial structure $\mathcal{S}' = ([i,j-1]_F.\sigma; \mathcal{P}; E_{L_S})$ on the stack R for later refinement.

Next we scan for all possible outermost pairs $l \cdot j$. If for a particular $l \cdot j$ the criterion

$$F^5_{l-1} + C_{l,j} + d^5_{l,j,l-1} + d^3_{l,j,j+1} + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \leq E_{\min} + \delta \tag{17}$$

is fulfilled, we push the refinement $\mathcal{S}' = ([l,j]_C.[i,l-1]_F.\sigma; \mathcal{P}; E_{L_S} + d^5_{l,j,l-1} + d^3_{l,j,j+1})$ on the stack R. Note that we do not to add the base pair $l \cdot j$ here, but we shall do so when refining the interval $[l,j]_C$ closed by it.

▶ case $E = C$ (backtrack in C)

Position i pairs with j in the popped segment $[i,j]_C$. We first take it to be a hairpin. If

$$\mathcal{H}(i,j) + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \leq E_{\min} + \delta, \tag{18}$$

we obtain a refinement of $\mathcal{S}$, $\mathcal{S}' = (\sigma; \mathcal{P} \cup \{i \cdot j\}; E_{L_S} + \mathcal{H}(i,j))$, which is pushed on the structure stack R. Next, we construct stacks, interior loops and bulges by scanning for all admissible pairs $p \cdot q$, and checking the condition

$$C_{i,j} + \mathcal{I}(i,j,p,q) + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \leq E_{\min} + \delta. \tag{19}$$

Each time inequality (19) is fulfilled, we obtain a refinement of $\mathcal{S}$, $\mathcal{S}' = ([p,q]_C.\sigma; \mathcal{P} \cup \{i \cdot j, p \cdot q\}; E_{L_S} + \mathcal{I}(i,j,p,q))$ which is stacked on R. We proceed to construct multiloops in correspondence with the third term of equation (11). To this end we loop over k, monitoring condition

$$\begin{aligned} F^M_{i+1,k} + F^{M1}_{k+1,j-1} + d^5_{i,j,i+1} + d^3_{i,j,j-1} + \mathcal{M}_C \\ + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \leq E_{\min} + \delta, \end{aligned} \tag{20}$$

yielding more $\mathcal{S}$-refinements, $\mathcal{S}' = ([k+1,j-1]_{M1}.[i+1,k]_M.\sigma; \mathcal{P} \cup \{i \cdot j\}; E_{L_S} + d^5_{i,j,i+1} + d^3_{i,j,j-1} + \mathcal{M}_C)$, to be pushed on the partial structure stack.

▶ case $E = M1$ (multiloop backtrack in F^{M1})

Equation (12) is effectively traced back by nibbling away at the $3'$ end, and checking for a base pair which initiates the stem of the F^{M1}-segment under consideration. We first eat away at the $3'$ end:

$$F^{M1}_{i,j-1} + \mathcal{M}_B + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \leq E_{\min} + \delta. \tag{21}$$

If (21) holds, we push $\mathcal{S}' = ([i,j-1]_{M1}.\sigma; \mathcal{P}; E_{L_S} + \mathcal{M}_B)$. When $\mathcal{S}'$ is popped again, the $3'$ nibbling will continue.

We next check whether i and j can pair. If they can, we must consider

$$C_{i,j} + d^5_{i,j,i-1} + d^3_{i,j,j+1} + \mathcal{M}_I + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \leq E_{\min} + \delta, \tag{22}$$

which leads us to push $\mathcal{S}' = ([i,j]_C.\sigma; \mathcal{P}; E_{L_S} + d^5_{i,j,i-1} + d^3_{i,j,j+1} + \mathcal{M}_I)$.

▶ case $E = M$ (multiloop backtrack in F^M)

To trace back equation (13), we insert the definition of F^{M1}, (12), into (13). As in the F^{M1} case, we start nibbling away at the $3'$-end, and also consider an interior base pair. This takes partially care of the F^{M1}-term in equation (13). The procedure here follows exactly the $E = M1$ case, except that the nibbled segment $[i,j-1]$, to be pushed, is now marked M.

To complete equation (13) we only need to loop over k, considering pairs $k+1\cdot j$ which fulfill

$$F^M_{i,k} + C_{k+1,j} + d^5_{k+1,j,k} + d^3_{k+1,j,j+1} + \mathcal{M}_I \\ + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \le E_{\min} + \delta. \tag{23}$$

The corresponding refinements $\mathcal{S}' = ([k+1,j]_C.[i,k]_M.\sigma; \mathcal{P}; E_{L_S} + d^5_{k+1,j,k} + d^3_{k+1,j,j+1} + \mathcal{M}_I)$ are pushed on R.

To cover the case in which the multiloop decomposition segment $[i,j]_M$ contains exactly one interior base pair, we complete the backtrack of equation (14) by looping over k, searching for pairs $k+1\cdot j$ such that

$$C_{k+1,j} + d^5_{k+1,j,k} + d^3_{k+1,j,j+1} + \mathcal{M}_I + (k-i+1)\cdot\mathcal{M}_B \\ + E_{L_S} + \sum_{[k,l]\in s} E_{k,l} \le E_{\min} + \delta, \tag{24}$$

and pushing $\mathcal{S}' = ([k+1,j]_C.\sigma; \mathcal{P}; E_{L_S} + d^5_{k+1,j,k} + d^3_{k+1,j,j+1} + \mathcal{M}_I + \mathcal{M}_B)\cdot(k-i+1)$.

▶ If $\mathcal{S} = ([i,j]_E.\sigma; \mathcal{P}; E_{L_S})$ caused no refinement to be pushed on R, then push $(\sigma; \mathcal{P}; E_{L_S})$.

C TRANSLATION OF MODIFIED NUCLEOTIDES

All tRNA sequences of *E.coli* are from the compilation of Steegborn [28] which can be obtained via anonymous ftp from EMBL Heidelberg, `ftp.embl-heidelberg.de`, in directory `/pub/databases/trna`.

Bases are translated as suggested by Higgs [22]. Most modified bases occur only in loop regions and are therefore classified as non-bonding. Only the following bases are often found in paired regions and are translated to their canonical equivalents:

```
H  (?A)        unknown modified adenosine
^  (Ar(p))     2'-O-ribosyladenosine (phosphat)
```

```
<  (?C)        unknown modified cytidine
B  (Cm)        2'-O-methylcytidine
M  (ac4C)      N4-acetylcytidine
?  (m5C)       5-methylcytidine

;  (G)         unknown modified guanosine
L  (m2G)       N2-methylguanosine
#  (Gm)        2'-O-methylguanosine
R  (m22G)      N2,N2-dimethylguanosine

N  (?U)        unknown modified uridine
J  (Um)        2'-O-methyluridine
P  (psi)       pseudouridine
]  (m1psi)     1-methylpseudouridine
Z  (psi m)     2'-O-methylpseudouridine
```

References

1. Nussinov, R., Piecznik, G., Griggs, J. R. & Kleitman, D. J. (1978) *SIAM J. Appl. Math.* **35**(1), 68–82.

2. Waterman, M. S. & Smith, T. F. (1978) *Mathematical Biosciences* **42**, 257–266.

3. Nussinov, R. & Jacobson, A. B. (1980) *Proc. Natl. Acad. Sci. USA* **77**(11), 6309–6313.

4. Zuker, M. & Stiegler, P. (1981) *Nucleic Acids Research* **9**, 133–148.

5. Zuker, M. & Sankoff, D. (1984) *Bull.Math.Biol.* **46**(4), 591–621.

6. Zuker, M. (1989) *Science* **244**, 48–52.

7. Nakaya, A., Yamamoto, K. & Yonezawa, A. (1995) *Comput. Applic. Biosci.* **11**(6), 685–692.

8. Nakaya, A., Yonezawa, A. & Yamamoto, K. (1996) *J. theor. Biol.* **183**, 105–117.

9. Waterman, M. S. & Byers, T. (1985) *Mathematical Biosciences* **77**, 179–188.

10. Waterman, M. S. (1995) *Introduction to Computational Biology: Sequences, Maps and Genomes.* Chapman & Hall, London.

11. McCaskill, J. S. (1990) *Biopolymers* **29**, 1105–1119.

12. Hofacker, I. L., Fontana, W., Stadler, P. F. & Schuster, P. (1994-1998) `http://www.tbi.univie.ac.at/~ivo/RNA/`, (Free Software).

13. Cupal, J., Hofacker, I. L. & Stadler, P. F. In *Computer Science and Biology 96 (Proceedings of the German Conference on Bioinformatics),* Hofstädt, R., Lengauer, T., Löffler, M. & Schomburg, D., editors, 184–186 (Univeristät Leipzig, Leipzig (Germany), 1996).

14. Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Muller, P., Mathews, D. H. & Zuker, M. (1994) *Proc. Natl. Acad. Sci.* **91**, 9218 – 9222.

15. Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986) *Proc.Natl.Acad.Sci.USA* **83**, 9373–9377.

16. Turner, D. H., Sugimoto, N. & Freier, S. (1988) *Annual Review of Biophysics and Biophysical Chemistry* **17**, 167–192.

17. Jaeger, J. A., Turner, D. H. & Zuker, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7706–7710.

18. He, L., Kierzek, R., SantaLucia, J., Walter, A. & Turner, D. (1991) *Biochemistry* **30**, 11124.

19. Ward, J. H. (1963) *Journal of the American Statistical Association* **58**, 236–244.

20. Ninio, J. (1979) *Biochimie* **61**, 1133.

21. Higgs, P. G. (1993) *J.Phys.I (France)* **3**, 43.

22. Higgs, P. G. (1995) *J.Chem.Soc.Faraday Trans.* **91**(16), 2531–2540.

23. Fontana, W., Konings, D. A. M., Stadler, P. F. & Schuster, P. (1993) *Biopolymers* **33**, 1389–1404.

24. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M. & Schuster, P. (1994) *Monatsh.Chem.* **125**(2), 167–188.

25. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. (1994) *Proc. Roy. Soc. (London) B* **255**, 279–284.

26. Huynen, M., Perelson, A., Vieira, W. & Stadler, P. (1996) *Journal of Computational Biology* **3**(2), 253–274.

27. Wuchty, S. Master's thesis, University of Vienna, (1998) .

28. Steegborn, C., Steinberg, S., Huebel, F. & Sprinzl, M. (1995) *Nucl. Acids Res.* **24**(1).

29. Shapiro, B. A. (1988) *CABIOS* **4**(3), 387–393.

30. Wuchty, S., Fontana, W. & Hofacker, I. L. (1998) `http://www.tbi.univie.ac.at/~ivo/RNA/`, (Free Software).